Supplementary Notes on Sampling Distributions
EPSY 5261

Now that you have a better understanding of the normal distribution, we will begin talking about another important distribution in statistics: the sampling distribution of the mean. This distribution is important because it lays the foundation for our ability to make inferences about populations based on samples.

The purpose of these notes is to try to clarify some of the issues you may find confusing as you read about sampling distributions. I also want to provide more examples for you and prepare you for your next homework assignment.

What is a sampling distribution?

We can create sampling distributions for any statistic (e.g., the mean, median, standard deviation, etc.). Our focus this semester will be on the **sampling distribution of the mean**. Your book refers to this as the **sampling distribution of possible sample means**.

Imagine you have a population of test scores that is normally distributed. Let's say this population has a mean of $\mu = 91.25$ and a standard deviation of $\sigma = 7.40$ (note I am now using the symbols μ and σ to refer to the mean and standard deviation; you might remember these symbols are POPULATION PARAMETERS). You are interested in taking <u>one</u> sample of size n = 20 from this population. You are taking this sample because you want to make an inference about the population.

How can we go about making an inference based on <u>one</u> sample? We need to have an understanding of how this ONE sample compares to the population it comes from and how it compares to all other possible samples we might have drawn at random from the population. In research, you will probably only have the resources to take one random sample, and we need a theory to allow us to determine how this one sample can allow us to make a claim or test a claim about a population.

The sample of size n = 20 that you draw is only one of many possible samples of size n = 20 that could have been drawn from this population. Let's say we had a way to simulate drawing **all possible samples** of size n = 20 from this population. If we take each sample, find the sample mean ($\bar{x}$), and then create a graph or a distribution of all of these sample means, we have what is called a **sampling distribution of the mean**.

Whenever we take all possible samples of a particular size (in this case, size n = 20) from a particular population, find the means of each of these samples, and then graph these means, we create what is known as the **sampling distribution of the mean**.

Characteristics of the sampling distribution

The sampling distribution of the mean will have a distinct shape, center, and spread (or variability), depending on what the original population looks like and how big the sample is.

- **Shape**: If the original population is NORMAL in shape, the sampling distribution of the mean will be approximately normal in shape as well, regardless of sample size. If the original population is not normal in shape (e.g., it is SKEWED, BIMODAL, TRIMODAL, etc.), the sampling distribution will only become normal in shape as we increase sample size. A "big enough" sample size to ensure a normal sampling distribution, when the population is not normal, has been found to be roughly about n = 30 (although this depends on just how "abnormal" the population is to begin with; generally, populations that are markedly skewed require larger sample sizes in order to ensure a normal sampling distribution).
- **Center**: The sampling distribution will have a **mean (or average)** equal to the population mean (μ). Note in some of your homework problems, you might be asked to find the "**mean of the sampling distribution of possible sample means**." This means that we want you to find the center, or mean, of the sampling distribution, and this is equal to the population mean.
- **Spread/Variability**: The sampling distribution will have a **standard deviation** equal to the population standard deviation divided by the square root of the sample size ($\frac{\sigma}{\sqrt{n}}$). In other words, the sampling distribution will be LESS VARIABLE than the original population, and it gets less and less variable as the sample size (n) increases. If you are ever asked to find the "**standard deviation of the sampling distribution of possible sample means**," this means you should use this formula to find the standard deviation of the sampling distribution.

In this particular example, we'd expect the shape of the sampling distribution to be NORMAL (since we are told the original population is normal), the mean of the sampling distribution to equal 91.25, and the standard deviation to equal $\frac{7.40}{\sqrt{20}} \approx 1.66$. The sampling distribution is thus centered where the population is centered but has less variability (e.g., a smaller standard deviation, or less spread) than the original population.

As you are thinking more about sampling distributions, please keep some important ideas in mind:

- Sampling distributions are THEORETICAL constructs. In practice, you would never go out and construct your own sampling distribution. To do so, you'd need to be able to take all possible samples of a certain size from the population you are interested in, and this is not practical. We know the properties of sampling distributions based on the simulations done by statisticians, and we use what is known about sampling distributions

to make inferences based on single samples. In practice, you will likely only take ONE sample from the population of interest, but as I said earlier, in order to make a valid inference about the population, you need to understand how your sample compares to the population from which it comes from AND to all other possible samples that could have been drawn from the same population.

- As you are first learning about sampling distributions, you will be told specific values of the population mean (μ) and the population standard deviation (σ). In reality, these parameters are generally not know and you use **sample statistics** such as the sample mean ($\bar{x}$) and the sample standard deviation (s) in order to make inferences about these unknown **population parameters**.
- Two important laws or theorems allow us to better understand how sampling distributions behave.
    - The **Law of Large Numbers** tells us that as we increase the size of our sample, the sample mean ($\bar{x}$) should get closer and closer to the population mean (μ). This makes intuitive sense when you think about it. If we take a larger "chunk" of the population (by taking a big random sample), we should obtain a sample mean that is a good representation of the population mean. If you imagined taking all possible random samples of a large size from a particular population and plotting/graphing all the sample means, you'd expect them all to cluster close to the actual value of μ. Basically, the Law of Large Numbers tells us how one sample compares to the population it comes from. Think about how the **Drawing Samples Lab** you worked on this week illustrates this law.
    - The **Central Limit Theorem** tells us something about what the sampling distribution will look like. It states that if the sample size (n) is sufficiently large (with "large" being defined as roughly about n = 30), the sampling distribution will be approximately normal in shape, with a mean = μ and a standard deviation = $\frac{\sigma}{\sqrt{n}}$. If the original population is not normal, the sampling distribution will not be normal either unless the sample size is large.
- There are several important symbols that are used in Chapter 9. It is important to distinguish among these symbols. I've tried to help you with this by putting together the following table. These are the symbols I want you to become most familiar with, for you will see them over and over again for the rest of the semester.

Symbols used to describe **Samples**	Symbols used to describe **Populations**	Symbols used to describe **Sampling Distributions**
Sample mean = $\bar{x}$	Population mean = μ	Mean = μ (Note: sometimes this mean is denoted as $\mu_{\bar{x}}$)
Sample standard deviation = s	Population standard deviation = σ	Standard deviation = $\frac{\sigma}{\sqrt{n}}$ (Note: sometimes this standard deviation is denoted as $\sigma_{\bar{x}}$)
		Standard error of the mean= $\frac{s}{\sqrt{n}}$ (Note: we use the standard error when we do not know σ and must therefore approximate σ based on s)

Applying knowledge of sampling distributions

This week in class, you will learn more about the sampling distribution by using a program called Sampling SIM. Our goal in having you work through the Sampling SIM program is to better understand the concept of sampling distributions. The Sampling SIM program allows you to simulate drawing many samples of particular sizes from particular populations. As you perform these simulations, we hope you will notice how factors such as the shape of the original population and the size of the sample affect the resulting sampling distributions. Remember, we expect that if the population is not normal, the sampling distribution will not be normal unless the sample size is large. We also expect that the center of the sampling distribution will be equal to (or close) to the population mean (μ) but the spread or variability of the sampling distribution will be smaller than the spread or variability of the population (since, in this case, our measure of spread is the standard deviation, and the standard deviation of the sampling distribution is equal to $\frac{\sigma}{\sqrt{n}}$). In the lab, you will not be creating true sampling distributions in that you will not be drawing **all possible samples** of a certain size from the populations. Instead, you will draw many, many samples so you can approximate the basic characteristics of the true sampling distribution.

Let's work through some applications of sampling distributions so we can see how we actually use this information to solve real problems.

The following population is described in a problem at the end of Chapter 9. I'm going to use this population information to illustrate some of the issues you should think about when trying to apply what you are learning about sampling distributions.